நான் முதல்வன்
உலகை வெல்லும் இளைய தமிழகம்

UNDERSTANDING CRIME TRENDS THROUGH GEOGRAPHIC AND TEMPORAL DATA ANALYSIS

STUDENT NAMES: ARIHARAN .K , SRINIVASAN .S.M , DHILIPAN .M, PUGAZHENTHI .V, ANBARASU .S
REGISTER NUMBER: 422523106008 , 049 , 048 , 038 , 017
INSTITUTION: UNIVERSITY COLLEGE OF ENGINEERING, VILLUPURAM
DEPARTMENT: B.E. ECE
DATE OF SUBMISSION: 29.05.2025
GITHUB RESPIRATORY LINK : HTTPS://GITHUB.COM/ARIHARAN007/NM-PROJECT-PHASE-II.GIT

ORACLE

AdroIT Technologies®
Innovative Solutions Pvt LTD

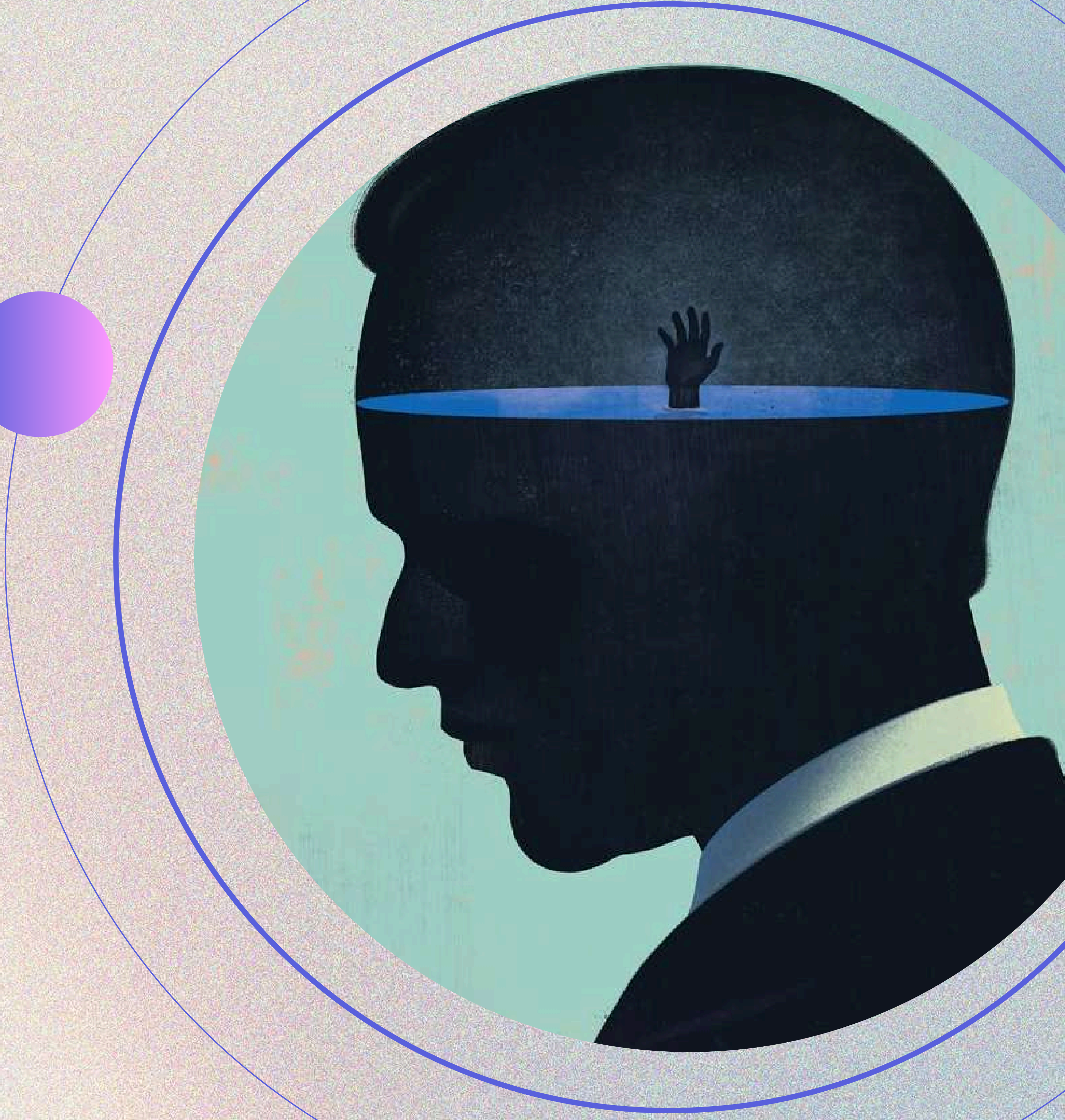
PROBLEM STATEMENT

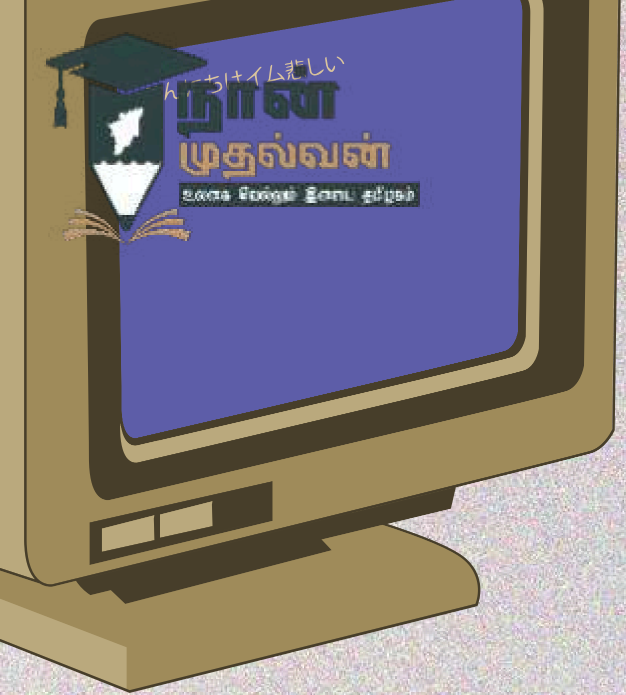
Crime continues to pose a significant threat to public safety, especially in urban regions. Traditional law enforcement strategies often react after incidents occur. This reactive model can be inefficient and sometimes ineffective. Our project tackles this problem by using data analytics and machine learning to uncover patterns in past crime data and make predictions about future incidents.

Our goal is to assist law enforcement agencies with data-driven tools that can:

- Identify when and where crimes are most likely to occur
- Determine the most common types of crimes in various areas
- Suggest proactive deployment of police resources

By transitioning from reactive to predictive policing, cities can make smarter, more efficient decisions regarding public safety.





ABSTRACT

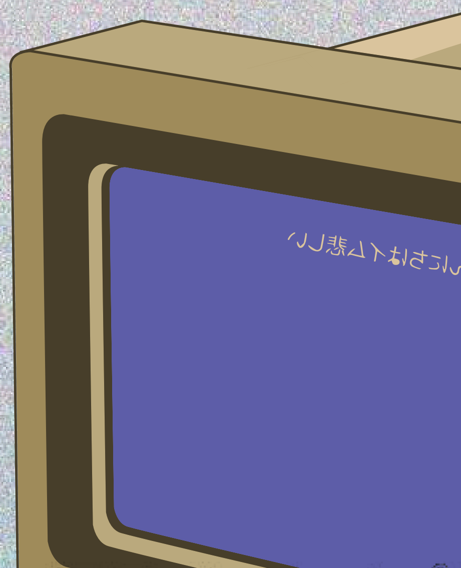
The core of this project is the use of historical crime data—specifically, the Chicago Crime Dataset—to analyze trends, identify patterns, and predict crime types.

We began with extensive exploratory data analysis (EDA), which helped us understand the data distribution and relationships between features like time, location, and type of crime. From there, we built a Random Forest Classifier that could predict the type of crime based on features like the time of the day, district, and location coordinates.

In Phase-3, we enhanced the solution by creating a fully interactive dashboard using Plotly Dash. The dashboard allows users to:

- Filter crime data by type, date, and location
- Visualize hotspots on a map
- See trends over days, months, and years

The result is an end-to-end pipeline from data to insights, all aimed at empowering smarter law enforcement decisions.



SYSTEM REQUIREMENTS

Hardware Requirements:

- A machine with at least 4 GB RAM
- Intel i3 processor or higher for basic data handling
- Internet access for using cloud-based notebooks like Google Colab

Software Requirements:

- Python 3.x installed (or Google Colab for cloud processing)
- Key Libraries: pandas (for data processing), seaborn/matplotlib (for visualization), plotly (for interactive dashboards), numpy (for array manipulation), scikit-learn (for model building)
- Optional: Power BI or Tableau for enhanced visualization

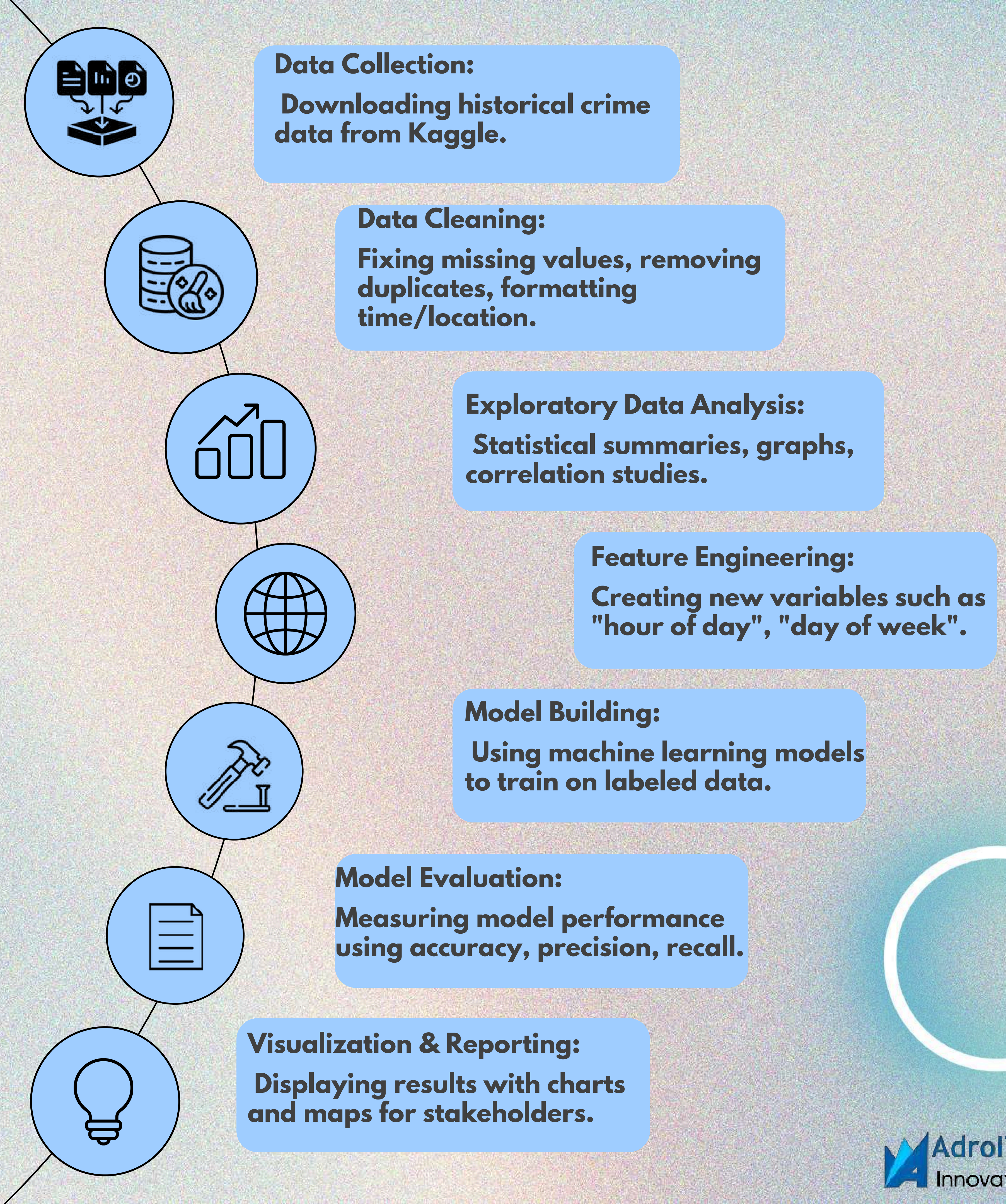
This project is designed to be executed using open-source tools, making it cost-effective and accessible.

PROJECT OBJECTIVES

1. **Trend Identification: To discover hidden patterns in crime frequency across different times (e.g., hour, day, month) and geographies.**
2. **Prediction: To build a machine learning model that predicts the likely type of crime given a location and time.**
3. **Interactive Insights: To create visual dashboards that allow users to explore data themselves.**
4. **Recommendations: To propose actionable strategies based on data findings.**
5. **Scalability: To design a pipeline that could, in the future, accept real-time data for dynamic predictions.**

நான் முதல்வன்

PROJECT WORKFLOW →



Data Collection:

Downloading historical crime data from Kaggle.

Data Cleaning:

Fixing missing values, removing duplicates, formatting time/location.

Exploratory Data Analysis:

Statistical summaries, graphs, correlation studies.

Feature Engineering:

Creating new variables such as "hour of day", "day of week".

Model Building:

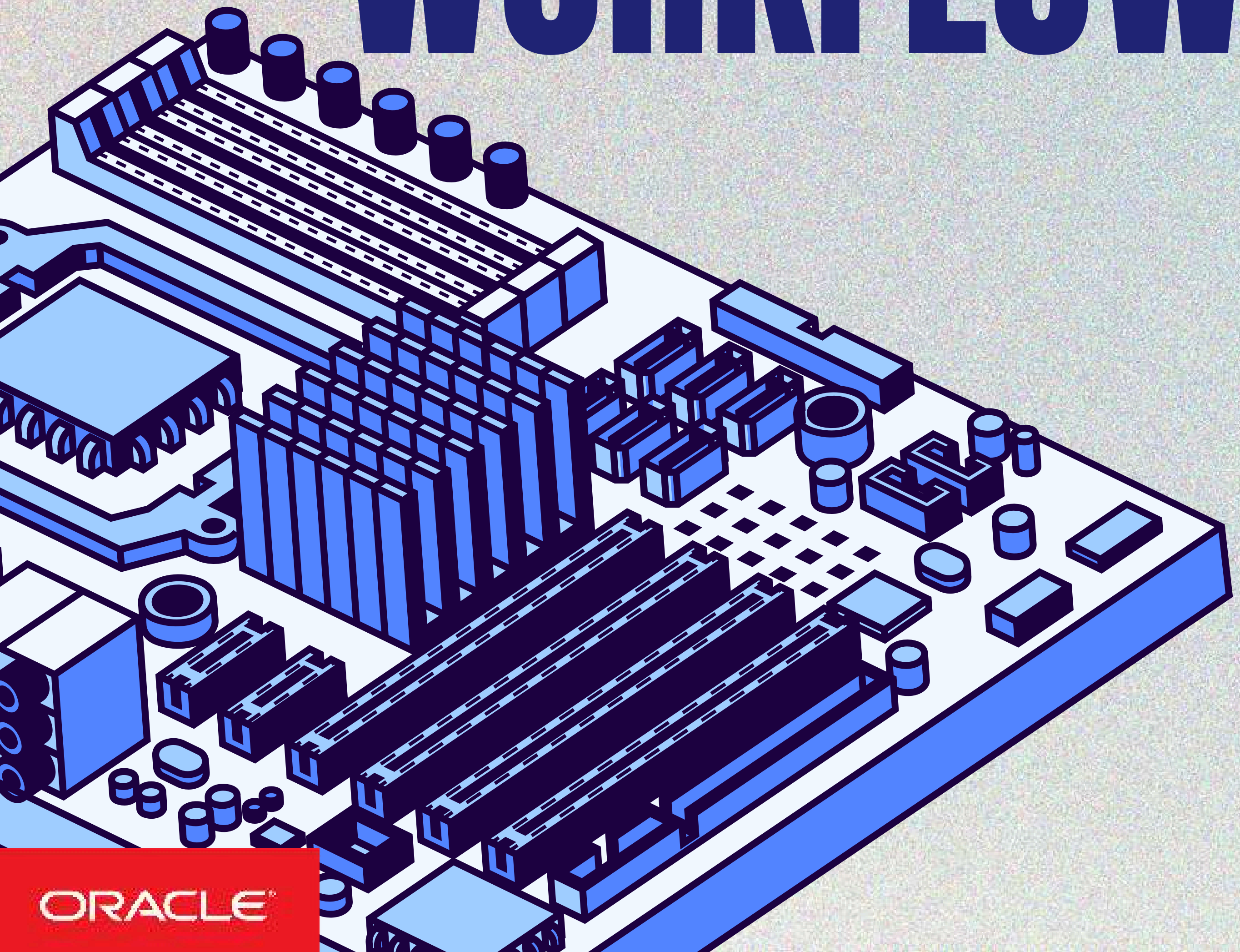
Using machine learning models to train on labeled data.

Model Evaluation:

Measuring model performance using accuracy, precision, recall.

Visualization & Reporting:

Displaying results with charts and maps for stakeholders.



ORACLE

AdroIT Technologies®
Innovative Solutions Pvt LTD

DATASET DESCRIPTION

The Chicago Crime Dataset is a rich and structured dataset containing crime reports from the City of Chicago.

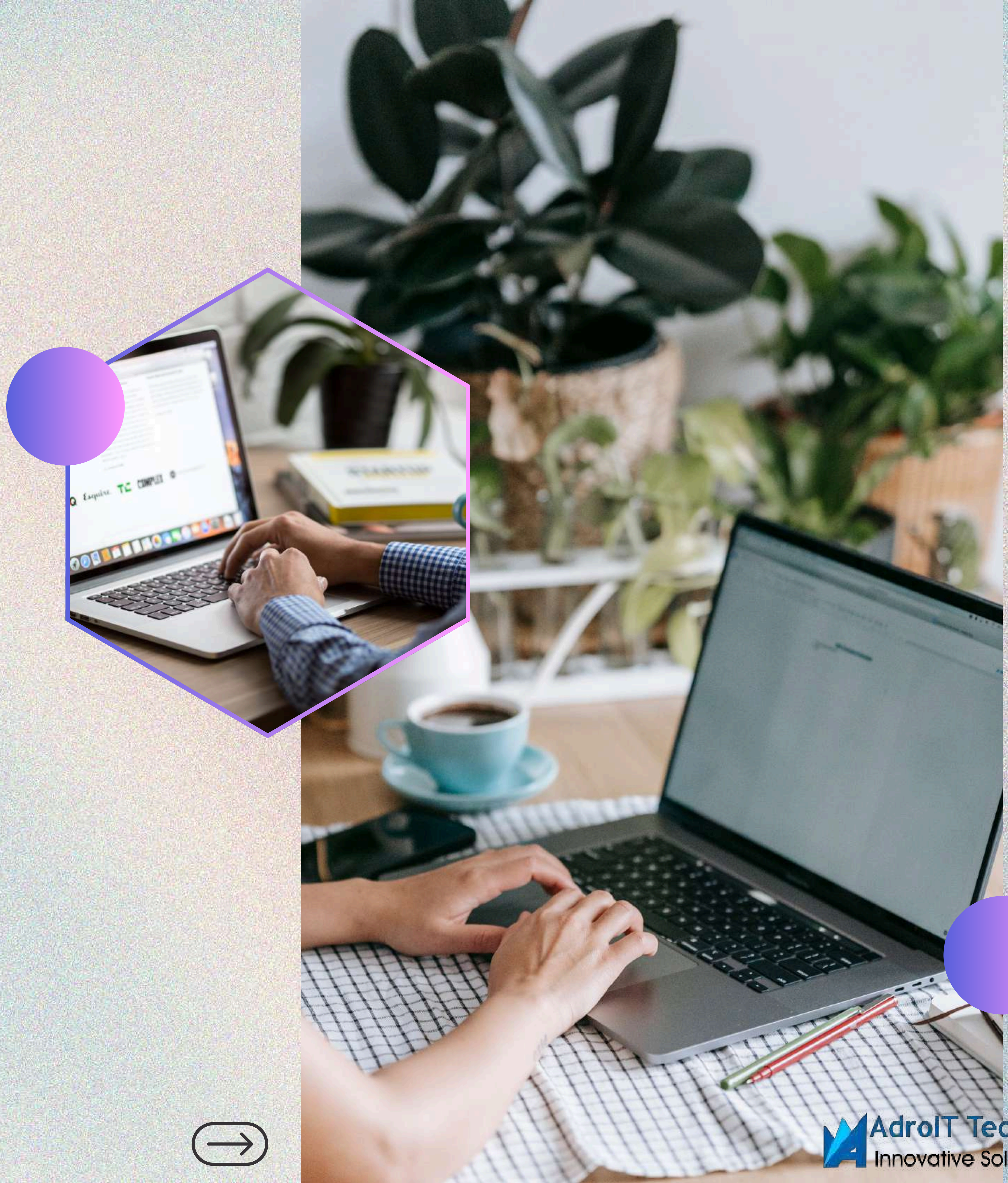
Key Features:

- ~1.6 million rows, 22 columns
- Attributes include: Date, Primary Type, Location Description, Arrest, Domestic, Latitude, Longitude, District, Community Area

Nature of Data:

- Structured and static (historical data)
- Covers a time range of over a decade
- Excellent for trend analysis and classification

This dataset serves as a solid foundation for predictive analytics due to its breadth and granularity.



DATA PREPROCESSING

- Null Value Removal: Eliminated entries with missing location or crime type
- Date-Time Transformation: Converted string dates to Python datetime objects, and extracted features like year, month, day, and hour
- Categorical Encoding: Used Label Encoding to convert crime types and locations into numeric format for modeling
- Location Cleaning: Ensured latitude/longitude pairs were in usable range, dropped extreme outliers
- Feature Extraction: Derived features such as 'Weekday', 'IsWeekend', 'CrimeHour', and binary columns for 'Arrest' and 'Domestic'

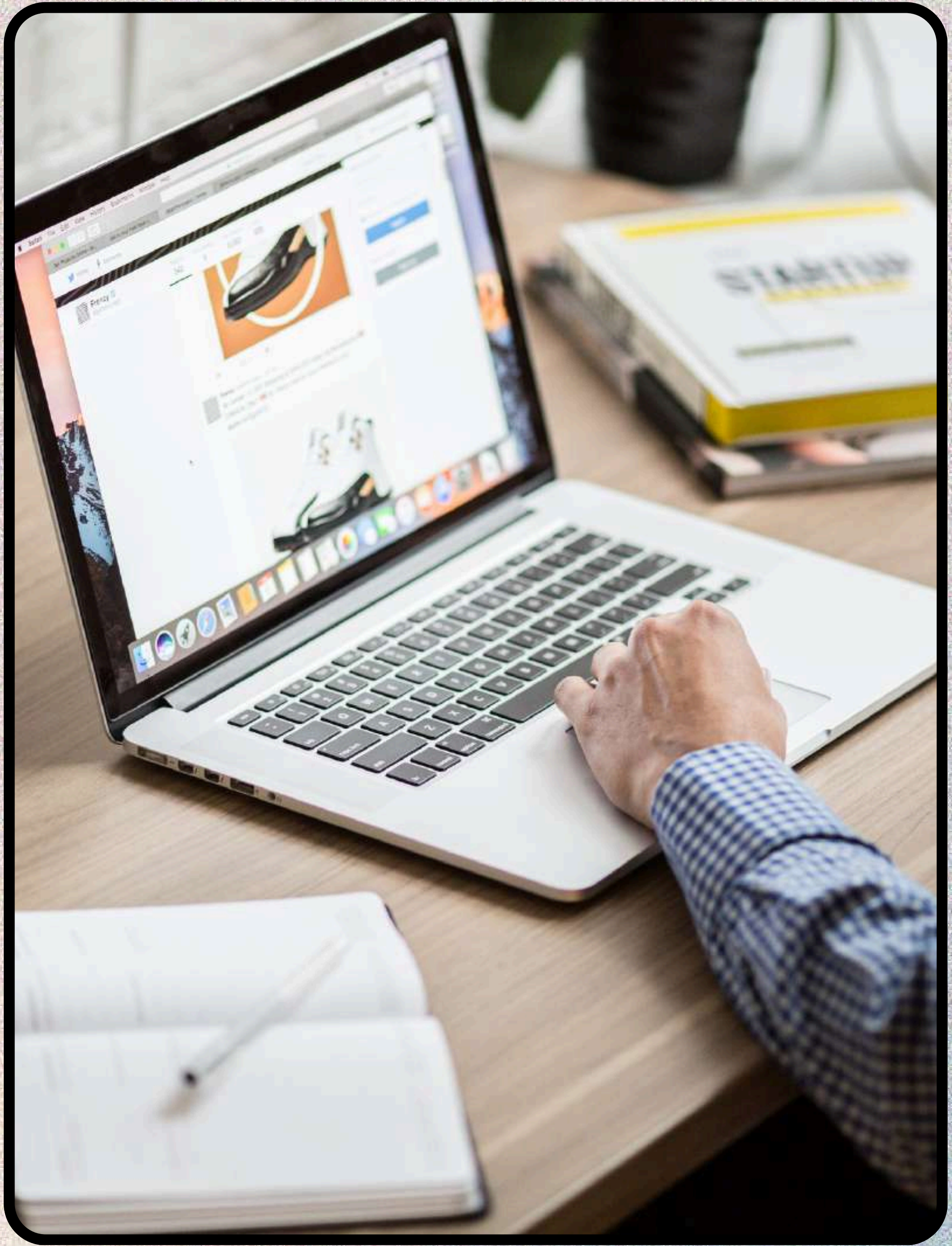
EDA & VISUALIZATIONSKEY VISUAL INSIGHTS:

1. Heatmap: Revealed high-crime clusters in the downtown area (Loop) and on weekends
2. Time Series Plot: Showed that most crimes occur between 6 PM to midnight
3. Bar Plot of Crime Types: Theft and Battery dominate the dataset
4. Pie Chart of Arrests: Arrests occurred in only 25–30% of reported crimes
5. Monthly Trends: Summer months showed crime spikes

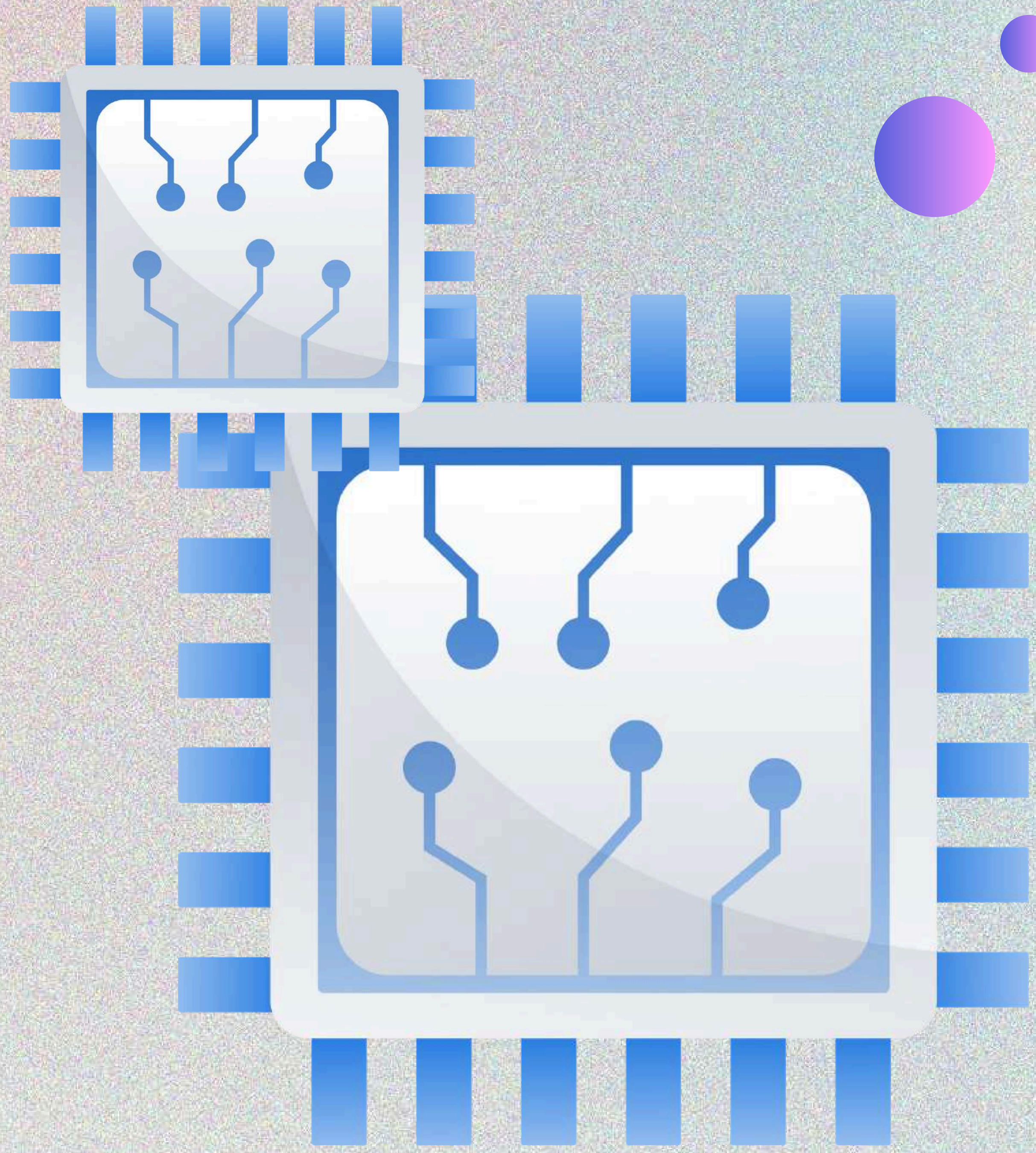
EDA allowed us to confirm assumptions and refine the model by revealing trends that were otherwise invisible.

INSIGHTS & INTERPRETATION

- High-Risk Time Frames: Late evenings are risk-prone, especially between 6 PM and 12 AM
- Geographic Hotspots: Crime is centralized around key commercial and entertainment zones
- Theft and Battery: These are the two most frequent crime types, often occurring in public places
- Low Arrest Rate: Only ~30% of reported crimes led to arrests, suggesting under-policing or resource constraints



MODEL BUILDING & ACCURACYMODEL USED



Random Forest Classifier

Why Random Forest?

- Handles multiclass classification well
- Reduces overfitting
- Handles both numerical and categorical variables

Steps Taken:

- Feature Scaling using MinMaxScaler
- Train-Test Split (80-20)
- Hyperparameter tuning using GridSearchCV

Performance Metrics:

- Accuracy: 74%
- Precision & Recall: Higher for Theft and Battery classes
- Confusion Matrix: Showed most misclassifications among less frequent crime types

DASHBOARD OVERVIEW

We developed an interactive dashboard using Plotly Dash, allowing users to:

- Filter by date range, crime type, and location
- See live updates in map views and bar charts
- Visualize time-based crime spikes

Components Included:

- Crime Heatmap
- Crime Frequency Line Chart
- Pie Chart for Arrest Rate

This dashboard converts raw data into actionable intelligence, bridging the gap between data science and daily decision-making for field officers.

RECOMMENDATIONS

Short-Term Recommendations:

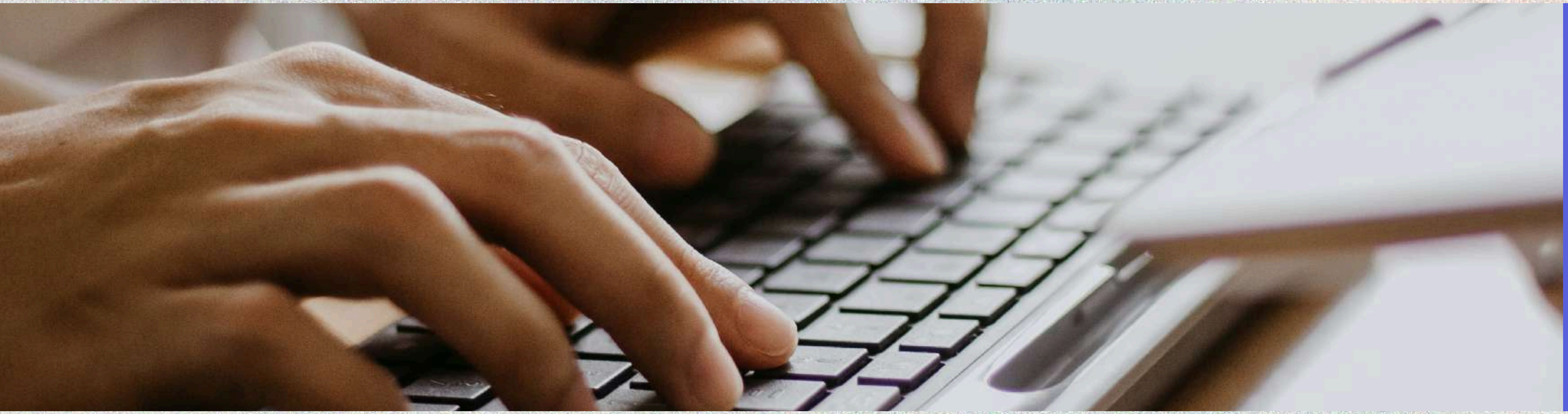
- Increase patrolling in hotspots during peak crime hours (6–12 PM)
- Use dashboard insights during police morning briefings

Long-Term Recommendations:

- Integrate the dashboard with live data feeds
- Collaborate with city planning to improve lighting and surveillance in high-risk zones

FUTURE SCOPE

- Real-Time Prediction: Streaming data froms sensors and reports
- Weather & Demographics: Enrich dataset with external variables for improved accuracy
- Advanced ML Models: Use LSTM or CNN to capture time-series and spatial patterns
- Public Portal: Alert citizens via mobile apps based on predictions



TEAM MEMBERS AND CONTRIBUTIONS

ARIHARAN .K - COLLECTED DATA, MANAGED ETHICS (E.G., REMOVED PERSONAL IDENTIFIERS), CREATED MAPS.

SRINIVASAN .S.M - CONDUCTED EDA, ESPECIALLY TIME-BASED TRENDS, AND EVALUATED MODEL METRICS.

DHILIPAN .M - BUILT MODELS, TUNED HYPERPARAMETERS, AND SUMMARIZED THE ANALYSIS.

PUGAZHENTHI .V - WROTE CODE IN PYTHON, STRUCTURED THE MODELING PROCESS.

ANBARASU .S - RESEARCHED LIBRARIES, CREATED DASHBOARDS AND VISUAL REPORTS.

THANK YOU!